

Worksheet 2 – Exploring Weka

Team: _____

Please place one finished copy of your work in your team folder

Save your work as I may ask you to demo it to the class.

1. Zoo Data

Draw the decision tree generated by Weka c4.5 (J4.8).

How accurate is C4.5 on this data? _____

What were the two classes that were confused the most? By that I mean an instance of class X was frequently mis-categorized as an instance of class Y

2. Breast Cancer Data

Here we will be comparing the performance of the ZeroR, 1R, C4.5 (J4.8) and Naïve Bayes.

ZeroR – what is the accuracy?

1R – what one rule is the best?

C4.5 – draw the decision tree

Compare the performance of these algorithms.

By default Weka uses 10-fold cross-validation.

Do you think changing this to 5 fold, or 20 fold, or 100 fold will have an effect on accuracy?

What's your prediction? What are the results of your experimentation?

3. Hypothyroid

Here I would like you to compare the performance of the algorithms (1R, C4.5, and Bayes) using training sets of different sizes. The original dataset has over 8,000 instances. Run it on that set then create three smaller datasets the smallest being around 100 instances. You might want to try the same thing with the mushroom data. Report on your findings. Is this what you expect?

With the original data file, in the “preprocess tab” click on Choose filter and select. filters → unsupervised → attribute → RemoveUseless. How many attributes were removed? (If you can find an easy way of finding what those attributes are let me know) Run a few of the classifiers again and compare the results.

4. Arabic documents

This is my data so it is not as Weka friendly as the other datasets. When you load a dataset you will need to remove the second attribute, filename. Then on the classify page change the attribute you are classifying on to the first one (source). It defaults to the last attribute. Again, in this one we are comparing the performance of the three algorithms. In this dataset we are trying to determine the city where an Arabic document was written. The other attributes are frequencies of common Arabic words. (The list of words is on the dataset page). I would like you to compare the performance of the algorithms as we vary the number of attributes. The original dataset has the frequencies of the 58 most common Arabic words. I have several other datasets including one with a 1,000 frequent words. So here describe the results. How do the algorithms compare? How does changing the number of attributes change things?

5. Spam, Spam, and Spam

Here we are looking at the Spam database.

What is the single best feature that discriminates Spam from non-Spam?

Compare the performance of the 3 algorithms

